AFRL-IF-RS-TR-2006-197
Final Technical Report
June 2006

EXPRESSIVE THREAT DETECTION VALIDATION FRAMEWORK

IET, Inc.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2006-197 has been reviewed and is approved for publication

APPROVED: /s/

DEBORAH A. CERINO
Project Engineer

FOR THE DIRECTOR: /s/

JOSEPH CAMERA, Chief
Information & Intelligence Exploitation Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JUNE 2006	Final	Apr 05 – Apr 06

4. TITLE AND SUBTITLE

EXPRESSIVE THREAT DETECTION VALIDATION FRAMEWORK

5a. CONTRACT NUMBER
FA8750-05-C-0092

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
31011G

6. AUTHOR(S)

Robert Schrag and Brandon Goldfedder

5d. PROJECT NUMBER
EAGL

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
IET, Inc.
1911 N. Fort Myer Drive, Suite 600
Arlington VA 22209

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/IFED
525 Brooks Rd
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2006-197

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; distribution unlimited. PA# 06-395

13. SUPPLEMENTARY NOTES
AFRL Project Engineer: Deborah A. Cerino, IFED, Deborah.Cerino@rl.af.mil

Under this project IET has developed a performance evaluation laboratory that consists of a synthetic dataset generator, a post-processor to load generated data into relational databases, and a hypothesis scorer. The objective of this effort was to generate synthetic data sets for the counter terrorism domain to support objective performance of existing threat detection tools/technology. The generator can generate a file with 100,000 individuals and 1,000,000 observable transactions in about 12 minutes. The generator has over 100 different parameters that may be varied to make the link discovery problem easier or harder in specific ways. This massively parameterized problem space supports flexible experimentation that can address the following kinds of questions:

" What problem characteristics most influence a given technology's performance?
" What is the observed capabilities envelope of a given technology across the parameterized problem space?
" What performance can we predict for a given technology on a dataset with given characteristics?

15. SUBJECT TERMS

Networked data, generating synthetic data, scoring networked entities

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Deborah Cerino
U	U	U	UL	49	19b. TELEPONE NUMBER (*Include area code*)

Abstract

We summarize the accomplishments of Information Extraction & Transport, Inc. (IET) in the performance of Government Contract Number FA8750-05-C-0092 to the Air Force Research Laboratory (AFRL). The purpose of IET's contract is to develop and deliver for Government use further capabilities in the performance evaluation laboratory (PE Lab) that IET developed earlier under the Air Force's Evidence Assessment, Grouping, Linking, and Evaluation (EAGLE) program.

IET has scaled up its pre-existing PE Lab to generate and score hypotheses from datasets that are ten times larger, delivered the scaled-up PE Lab with comprehensive documentation (including about 100 pages of new documentation), generated and delivered scaled-up datasets, participated in program-level activities to help define follow-on research, and developed detailed recommendations for future PE Lab development and PE Lab-based research.

Table of Contents

List of Figures

1 SUMMARY

We summarize the accomplishments of Information Extraction & Transport, Inc. (IET) in the performance of Government Contract Number FA8750-05-C-0092 to the Air Force Research Laboratory (AFRL). The purpose of IET's contract is to develop and deliver for Government use further capabilities in the performance evaluation laboratory (PE lab) that IET developed earlier under the Evidence Assessment, Grouping, Linking, and Evaluation (EAGLE) program.

IET has scaled up its pre-existing PE Lab to generate and score hypotheses from datasets that are ten times larger, delivered the scaled-up PE Lab with comprehensive documentation (including about 100 pages of new documentation), generated and delivered scaled-up datasets, participated in program-level activities to help define follow-on research, and developed detailed recommendations for future PE Lab development and PE Lab-based research.

We summarize our main conclusions as follows:

- PE Lab can naturally and productively support a Tangram Concept of Operations where threat hypothesis generation is based on the blackboard model (as it has been during EAGLE's technology integration experiments—TIEs).
- A next-generation PE Lab should include the following capabilities, for which we have developed initial designs:
    - Dynamic social networks
    - Explicit intelligence collection models
    - Temporal representation and reasoning foundations for evidence and ground truth

These conclusions are covered in greater detail in Section 6.

2 INTRODUCTION

The PE Lab consists of a synthetic dataset generator, a post-processor to load generated data into relational databases (DBs), and a hypothesis scorer. Included in the system concept, but not in the PE Lab proper, is a given threat detection component under evaluation. Such a component is presumed to employ link discovery (LD) technology—and is referred to herein as an LD component.

The PE Lab includes key components summarized in Figure 1, where the following graphical conventions hold:

- Square-cornered boxes represent artifacts.
- Round-cornered boxes represent components. The LD component is rendered three-dimensionally to highlight its status outside of the PE Lab proper.
- Solid arrows represent flow of artifacts.

- Dotted arrows represent the flow of control information.

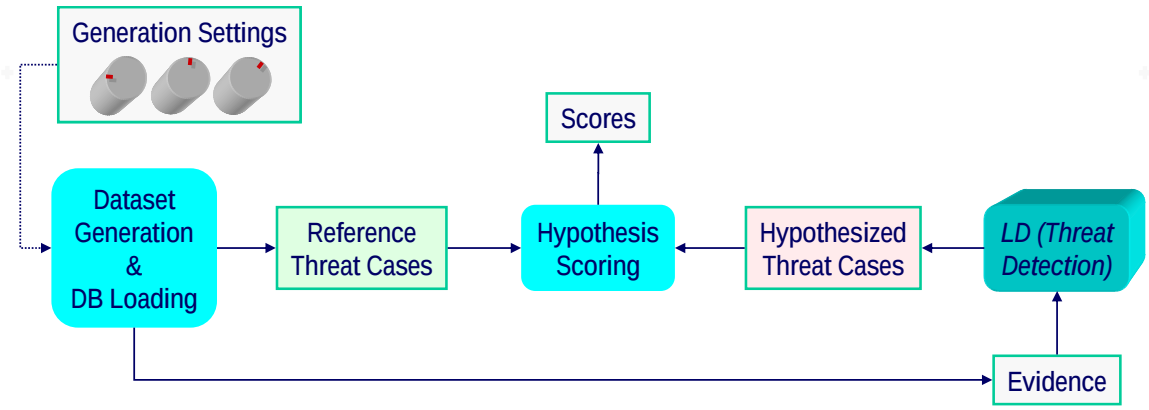

Figure 1: PE Lab Schematic

Dataset specifications (generator parameter settings) control creation of reference (ground-truth) threat cases for scoring and mixed threat and non-threat evidence for processing by threat detection. Detection examines evidence to detect threat phenomena, returning hypotheses for scoring. Scoring compares LD hypotheses to ground truth to yield scores.

IET initially developed the PE Lab under an earlier contract in the EAGLE program. Among its major accomplishments under the present contract, IET:

- Scaled up the EAGLE Year 3 (Y3) PE Lab that existed prior to contract start to support previously planned EAGLE Year 4 (Y4) levels for dataset size, population size, and signal-to-noise ratio (Section 3.1). This included updates for both dataset generation and hypothesis scoring.
- Delivered the scaled-up PE Lab, with comprehensive documentation (Section 3.2).
- Specified (Section 3.3), generated, validated, and delivered 41 Y4-compliant datasets.
- Participated in Tangram seedling activities, particularly by generating datasets to support experimentation and by analyzing experimental results.
- Developed detailed recommendations for:
    - PE Lab use in a Tangram concept of operations (Section 6.1).
    - Design of a next-generation PE Lab (Section 6.2) with:
        - Dynamic social networks.
        - Explicit intelligence collection models.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

Subsections provide additional information on the following topics:

- PE Lab documentation (Section 3.2)
- Specification of delivered datasets (Section 3.3)

3.1 Scaling up to Meet Y4 Requirements

IET optimized execution of the PE Lab's dataset generator in order to meet requirements for larger datasets. The generator is now 88% faster on datasets of "Y3-compliant" size (the largest size used in the most recent "Year 3" evaluation), generating the comma-separated value (CSV) flat files for a dataset with 100,000 individuals and 1,000,000 observable transactions in about 12 minutes on a new 32-bit Windows platform with 2 gigabytes running Franz Allegro Common Lisp 7.0. A dataset with 200,000 individuals and 2,000,000 observable transactions now takes about 30 minutes.

To generate datasets of Y4-compliant size, we procured a 64-bit Windows platform with 12 gigabytes of memory running 64-bit Franz Allegro Common Lisp. In this configuration, a dataset with 1,000,000 individuals and 10,000,000 observable transactions now takes about 4 hours to generate.

3.2 PE Lab Documentation

IET has developed and delivered the following comprehensive documentation for the Y4 PE Lab:

- Software Users Manual
- Software Product Specification
- Dataset Generator Guide
- Database Loader Guide
- Scoring Concepts Guide
- Scoring Software Guide
- Database Schema Specification

Much of this documentation—covering about 100 pages—is new with the Y4 PE Lab. The rest has been updated to reflect Y4 PE Lab capabilities.

Prior to the present contract, versions of the dataset generator and scoring documents were the main documentation released to participants. The new Dataset Generator Guide reflects Y4 capabilities (*e.g.,* cross-parameter constraints) and deletes material that is no longer relevant (*e.g.,* the tell model). We separated the old scoring document into the new Scoring Concepts and Scoring Software Guides. The Scoring Concepts Guide completely reorganizes and reformulates hypothesis scoring concepts to introduce them progressively and make them clear. The Scoring Software Guide has been updated to use terminology and notation consistent with the Scoring Concepts Guide and to reflect new scoring capabilities (*e.g.,* sparse matrix representation).

The new Database Schema Specification integrates conceptual material that earlier was available only in a slide presentation.

The Database Loader Guide is new.

The new Software Users Manual provides a CONOPS-level view of datasets' primary and secondary evidence and explains the contributions of the dataset generator and the database loader in establishing the time-dependent primary evidence database context. The Software Users Manual also includes a guide for the first-time user of the MySQL databases that we distribute with each dataset. The Software Users Manual also serves as a conceptual binder over all the other PE Lab documents (so that little material is repeated between documents). Virtually every aspect of the PE Lab software is now covered by the Software Users Manual and the other documents it binds.

3.3 Specification of Delivered Datasets

IET has used the following approach to develop a dataset mix (set of specifications) for the challenge problem datasets delivered under the contract:

1. Use "Y4-compliant"[1] settings established earlier in the program for noise/clutter, dataset size, and population size parameters.
2. Identify cross-parameter constraints concerning the Y4-compliant settings.
3. Specify, for each parameter, a probability distribution over its settings.
4. Determine the number *N* of datasets feasible to generate.
5. Generate a feasible dataset mix as follows:
    a. Generate a random dataset specification according to the probability distributions.
    b. Discard the specification if it does not satisfy the constraints, otherwise collect it.
    c. Repeat until *N* satisfactory specifications have been collected.
6. Diversify specifications in the developed feasible mix by swapping datasets' settings for a given parameter when constraints permit and this operation reduces the numbers of parameter-setting combinations shared across datasets.

[1] Y4-compliant settings were established earlier in the EAGLE program to reflect Y4 performance goals.

Figure 2 depicts the Y4-compliant parameter settings, giving other program years' parameter settings for context.

Population Size	Y1	Y2.5	Y3	Y4
Number of individuals	~1,000	~10,000	~100,000	~1000000
Mean threat group membership	20	80	80	80
Dev. threat group membership	5	20	20	20
Number of capabilities	50	100	150	200
Number of resources	50	100	150	200

Dataset Size	Y1	Y2.5	Y3	Y4
Number of observable transactions	~20000	~100,000	~1,000,000	~10000000

Noise, Clutter	Y1	Y2.5	Y3	Y4
Threat-to-clutter event ratio	0.08	0.008	0.0008	0.00008
Structured event SNR	0.08	0.008	0.0008	0.00008
Transaction event SNR	0.08	0.008	0.0008	0.00008
Individual SNR	0.4	0.08	0.008	0.0008
Group SNR	0.8	0.16	0.016	0.0016

Figure 2: Y4-compliant parameter settings

Figure 3 depicts the identified cross-parameter constraints.

Noise, Clutter	Dataset Size	Population Size	Pattern Complexity	Observability	Corruption	Entity Equivalence	Individual Duty Level	Rationale for prohibited combinations of settings listed in rows
	Y3	Y4						*People doing too few things during a simulation*
	Y2.5	Y4						
	Y1	Y4						
	Y2.5	Y3						
	Y1	Y3						
	Y1	Y2.5						
Y4	Y3							*Too few threat events (maybe none)*
Y4	Y2.5							
Y4	Y1							
Y3	Y2.5							
Y3	Y1							
Y2.5	Y1							
Y4	Y4		Fat					
Y3	Y3		Fat					
Y2.5	Y2.5		Fat					
	Y4	Y3	Fat				Easy	*Too many time ticks for incremental threat detection to be viable (currently)*
	Y4	Y2.5	Fat				Easy	
	Y4	Y1	Fat				Easy	
	Y3	Y2.5	Fat				Easy	
	Y3	Y1	Fat				Easy	
	Y2.5	Y1	Fat				Easy	
	Y4	Y3	Fat				Fair	
	Y4	Y2.5	Fat				Fair	
	Y4	Y1	Fat				Fair	
	Y3	Y2.5	Fat				Fair	
	Y3	Y1	Fat				Fair	
	Y2.5	Y1	Fat				Fair	
				Y2.5	Fair			*Do not adhere to consistent observability modeling*
				Covert	Fair			
				Y2.5	Hard			
				Covert	Hard			
Y3	Y4					Easy		*Too many individual equivalence classes to score*
Y3	Y4					Fair		
Y3	Y4					Hard		
Y2.5	Y4					Easy		
Y2.5	Y4					Fair		
Y2.5	Y4					Hard		
Y1	Y4					Easy		
Y1	Y4					Fair		
Y1	Y4					Hard		
Y1	Y3					Easy		
Y1	Y3					Fair		
Y1	Y3					Hard		
Y1	Y4					None		*Too many threat phenomena to score*

Figure 3: Cross-parameter constraints

Figure 4 illustrates weights used in the probability distribution for each parameter.

Group Connectivity	Noise, Clutter	Dataset Size	Population Size	Pattern Complexity	Observability	Y2.5 Raw Transaction Observability	Corruption	Entity Equivalence	Distinguishability (tell strength)	Target Duty Level	Individual Duty Level
None	Y1	Y1	Y1	Thin	Perfect	0.875	None	None	Easy	Easy	Easy
Easy	Y2.5	Y2.5	Y2.5	Fat	Easy	0.75	None	None	Easy	Fair	Fair
Fair	Y3	Y2.5	Y3		Fair	0.5	None	None	Easy	Hard	Hard
Hard	Y4	Y3	Y4		Fair		None	None	Easy		
		Y3	Y4		Hard		Fair	None	Fair		
		Y3	Y4		Hard		Fair	None	Fair		
		Y3			Y2.5		Fair	None	Hard		
		Y4			Covert		Hard	None			
		Y4					Hard	None			
		Y4						None			
		Y4						None			
		Y4						None			
		Y4						None			
		Y4						None			
		Y4						None			
		Y4						None			
		Y4						None			
		Y4						Easy			
		Y4						Easy			
		Y4						Easy			
		Y4						Easy			
		Y4						Fair			
		Y4						Fair			
		Y4						Hard			

Figure 4: Settings weights

In Figure 4, the number of settings tokens appearing under a dataset dimension indicates that setting's weight in a probability distribution used for setting selection.

Figure 5 represents 41 datasets resulting from the overall specification procedure (including the final diversification step).

Dataset Name	Group Connectivity	Noise, Clutter	Dataset Size	Population Size	Pattern Complexity	Observability	Y2.5 Raw Transaction Observability	Corruption	Entity Equivalence	Distinguishability (tell strength)	Target Duty Level	Individual Duty Level
EAGLE_Y4_3969	None	Y1	Y2.5	Y2.5	Thin	Fair		Fair	Easy	Easy	Easy	Easy
EAGLE_Y4_3971	None	Y4	Y4	Y2.5	Thin	Hard		None	None	Easy	Fair	Hard
EAGLE_Y4_3972	Fair	Y2.5	Y3	Y1	Thin	Hard		None	None	Easy	Hard	Fair
EAGLE_Y4_3973	Hard	Y3	Y4	Y4	Thin	Fair		None	None	Easy	Hard	Fair
EAGLE_Y4_3974	Fair	Y2.5	Y4	Y2.5	Thin	Covert	0.875	None	None	Easy	Easy	Easy
EAGLE_Y4_3975	Hard	Y3	Y4	Y4	Fat	Hard		Hard	None	Hard	Fair	Hard
EAGLE_Y4_3976	Fair	Y3	Y4	Y4	Thin	Hard		Fair	None	Easy	Easy	Hard
EAGLE_Y4_3977	Easy	Y3	Y4	Y2.5	Thin	Hard		None	Fair	Easy	Hard	Easy
EAGLE_Y4_3979	Hard	Y2.5	Y4	Y2.5	Thin	Easy		Fair	None	Easy	Hard	Hard
EAGLE_Y4_3980	Hard	Y2.5	Y3	Y2.5	Thin	Fair		None	Easy	Easy	Fair	Hard
EAGLE_Y4_3981	Hard	Y3	Y4	Y2.5	Fat	Fair		None	None	Easy	Easy	Hard
EAGLE_Y4_3982	Fair	Y2.5	Y4	Y4	Fat	Perfect		None	None	Easy	Fair	Fair
EAGLE_Y4_3983	None	Y2.5	Y4	Y2.5	Thin	Fair		None	None	Hard	Hard	Fair
EAGLE_Y4_3984	Hard	Y2.5	Y3	Y3	Fat	Fair		Hard	None	Easy	Hard	Easy
EAGLE_Y4_3985	Easy	Y3	Y4	Y4	Fat	Hard		Fair	None	Fair	Easy	Fair
EAGLE_Y4_3986	Easy	Y2.5	Y4	Y2.5	Fat	Y2.5	0.875	None	None	Fair	Hard	Hard
EAGLE_Y4_3988	Easy	Y3	Y4	Y4	Thin	Fair		Hard	None	Hard	Easy	Easy
EAGLE_Y4_3989	None	Y3	Y4	Y4	Fat	Covert	0.75	None	None	Easy	Hard	Hard
EAGLE_Y4_3993	None	Y4	Y4	Y4	Thin	Easy		None	None	Fair	Hard	Easy
EAGLE_Y4_3994	Fair	Y1	Y3	Y2.5	Thin	Fair		None	Easy	Easy	Hard	Easy
EAGLE_Y4_3995	None	Y2.5	Y4	Y4	Fat	Hard		None	None	Fair	Easy	Easy
EAGLE_Y4_3996	Easy	Y3	Y4	Y3	Thin	Covert	0.5	None	None	Fair	Fair	Easy
EAGLE_Y4_3997	None	Y2.5	Y4	Y4	Fat	Hard		None	None	Fair	Hard	Hard
EAGLE_Y4_3999	Hard	Y1	Y2.5	Y2.5	Fat	Hard		Hard	None	Easy	Fair	Hard
EAGLE_Y4_4000	None	Y2.5	Y4	Y3	Thin	Fair		None	None	Easy	Easy	Easy
EAGLE_Y4_4001	Easy	Y3	Y4	Y4	Thin	Fair		None	None	Easy	Fair	Hard
EAGLE_Y4_4002	Hard	Y4	Y4	Y4	Thin	Covert	0.875	None	None	Hard	Easy	Easy
EAGLE_Y4_4003	Hard	Y2.5	Y4	Y4	Fat	Fair		None	None	Hard	Fair	Easy
EAGLE_Y4_4004	Hard	Y2.5	Y4	Y4	Thin	Y2.5	0.75	None	None	Easy	Fair	Hard
EAGLE_Y4_4005	Hard	Y3	Y4	Y4	Thin	Fair		Hard	None	Easy	Easy	Fair
EAGLE_Y4_4006	None	Y4	Y4	Y1	Thin	Hard		Fair	Fair	Easy	Hard	Hard
EAGLE_Y4_4007	Fair	Y3	Y3	Y3	Thin	Fair		Hard	None	Easy	Easy	Easy
EAGLE_Y4_4008	Fair	Y3	Y4	Y2.5	Thin	Hard		Hard	Hard	Easy	Fair	Easy
EAGLE_Y4_4009	Hard	Y4	Y4	Y3	Thin	Hard		None	None	Easy	Fair	Fair
EAGLE_Y4_4010	Hard	Y3	Y4	Y4	Fat	Fair		Fair	None	Easy	Hard	Hard
EAGLE_Y4_4012	Easy	Y4	Y4	Y1	Thin	Hard		Hard	Fair	Fair	Hard	Easy
EAGLE_Y4_4013	Easy	Y3	Y4	Y4	Thin	Perfect		None	None	Easy	Fair	Hard
EAGLE_Y4_4015	Fair	Y2.5	Y4	Y4	Fat	Fair		None	None	Hard	Easy	Easy
EAGLE_Y4_4016	None	Y4	Y4	Y4	Thin	Hard		None	None	Fair	Fair	Fair
EAGLE_Y4_4017	Fair	Y2.5	Y3	Y2.5	Fat	Easy		None	None	Hard	Hard	Hard
EAGLE_Y4_4018	Easy	Y3	Y4	Y3	Fat	Hard		None	None	Fair	Fair	Hard

Figure 5: Candidate dataset mix

4 RESULTS AND DISCUSSION

The contract's main results are the scaled-up PE Lab software, comprehensive documentation, and scaled-up datasets that we have delivered (as summarized in Section 3.3). We also have developed detailed recommendations (presented in Section 6) for use of the PE Lab in a Tangram Concept of Operations and for design of a next-generation PE Lab to support future unclassified threat detection research.

5 CONCLUSIONS

We summarize our main conclusions as follows:

- PE Lab can naturally and productively support a Tangram Concept of Operations where threat hypothesis generation is based on the blackboard model (as it has been during EAGLE's technology integration experiments—TIEs).
- A next-generation PE Lab should include the following capabilities, for which we have developed initial designs:
    - Dynamic social networks
    - Explicit intelligence collection models
    - Temporal representation and reasoning foundations for evidence and ground truth

These conclusions are covered in greater detail in Section 6.

6 RECOMMENDATIONS

IET has developed recommendations in the following two main categories:

- Use of the PE Lab in a Tangram Concept of Operations (Section 6.1)
- Design of a next-generation PE Lab to support future unclassified threat detection research (Section 6.2)

6.1 Use of the PE Lab in a Tangram Concept of Operations (CONOPS)

Figure 6 depicts IET's Tangram vision of PE Lab-based dataset generation and scoring.

Generation parameter settings control creation of ground truth regarding the artificial world's resident entities, including threat individuals and groups. Evidence generation imposes partial observability and corruption (according to parameter settings) to create evidence from ground truth. Detection uses published and/or learned patterns and evidence to hypothesize ground truth phenomena not explicit in evidence. Hypothesis scoring compares hypothesized phenomena to ground truth cases to yield hypothesis quality metrics. In Tangram, CONOPS objectives will determine the kind(s) of hypotheses sought, the weights and costs to be used in scoring attributes of these hypotheses, and the overall objective function to apply to the various performance metrics, including those for hypothesis quality and for workflow execution (**Figure 7**).

We expect different operational circumstances to lead to different requirements regarding hypothesis quality (*e.g.*, completeness, accuracy, or certainty) and detection processing (*e.g.*, timeliness). Tangram challenge problem datasets should include plausible CONOPS objectives.

Figure 8 fleshes out metrics to be associated with PE Lab-generated datasets and their processing by Tangram threat detection components.

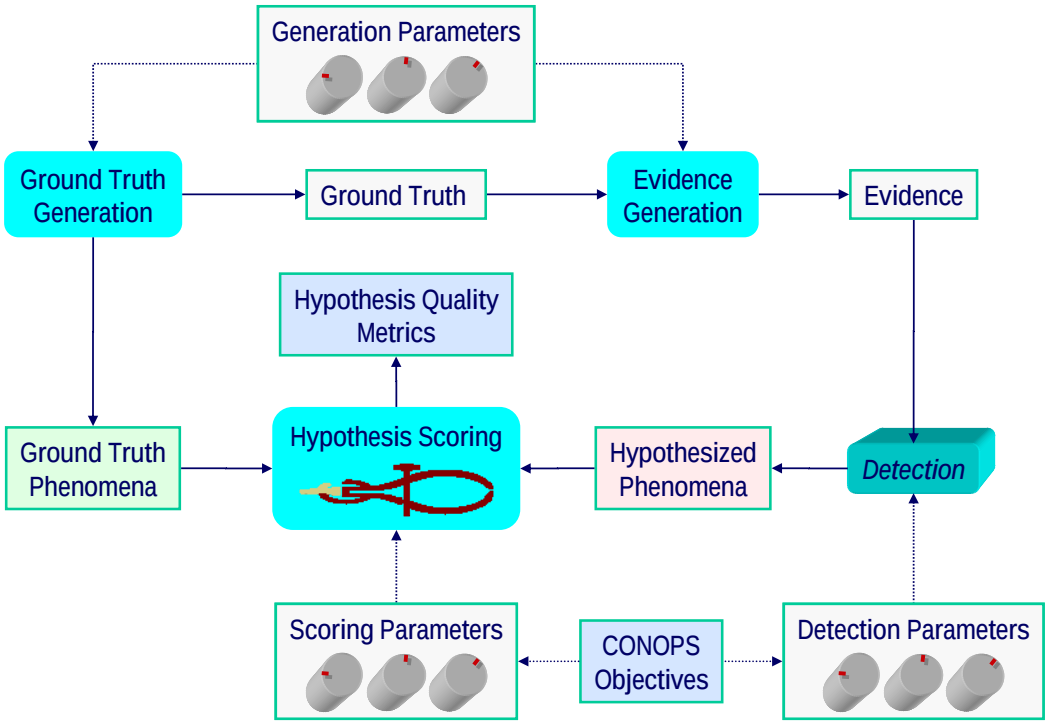

Figure 6: PE Lab-based dataset generation and scoring, in the Tangram context

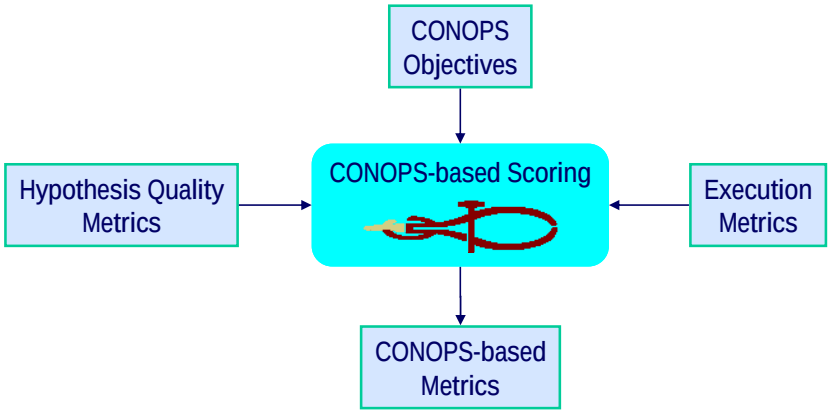

Figure 7: CONOPS-based scoring

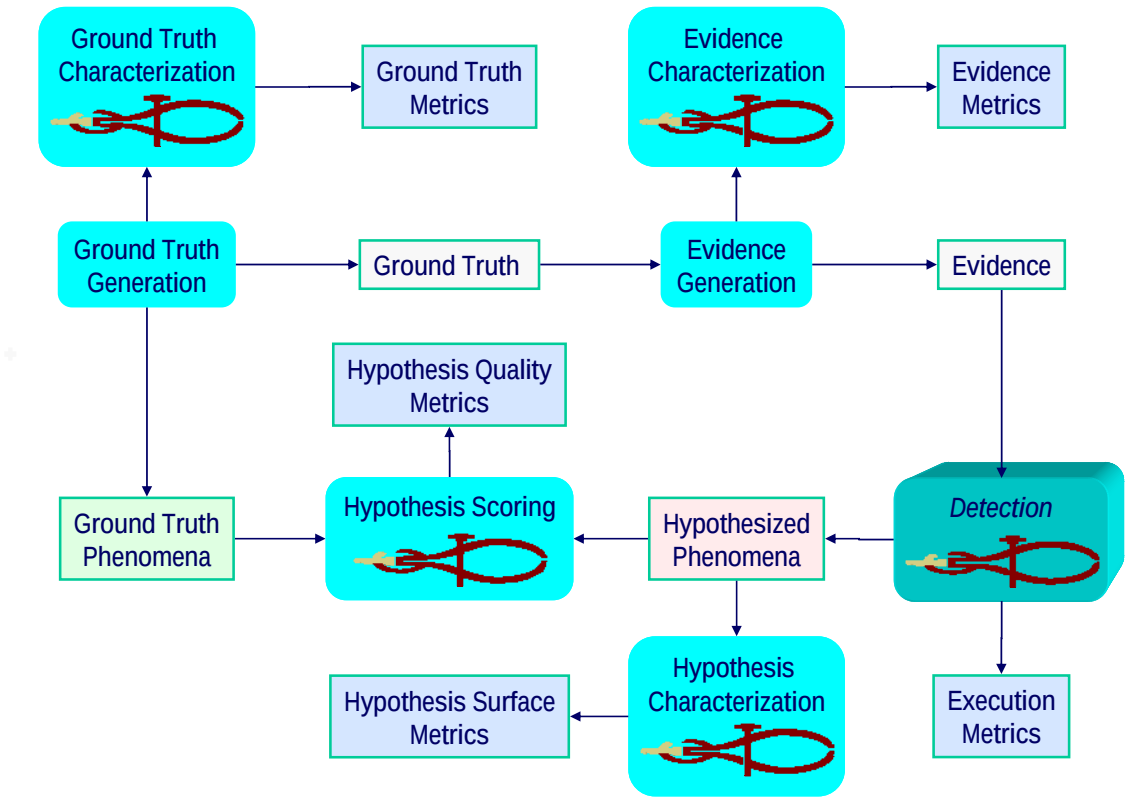

Figure 8: Tangram-relevant metrics, in the PE Lab context

Ground truth metrics, while not tenable in an operational context (where no ground truth is available), nonetheless may afford diagnostic value and are easily computed (*via* evidence characterization) because the PE Lab's evidence and ground truth follow the same schema.

In the workflow developed in response to a given top-level Tangram processing requirement, the "detection" process denoted in Figure 6 and Figure 8 may be unpacked into individual lower-level processing requirements that may be handled by different components. The key challenge for a Tangram system is to develop an effective workflow by selecting appropriate processing components, input/output relationships, and component parameter settings at appropriate times. We expect this selection to be based (at least in part) on historical performance information collected over many trials that pit different workflows against different, diverse datasets.

Figure 9 exhibits different kinds of historical information we expect to be recorded for these trials.

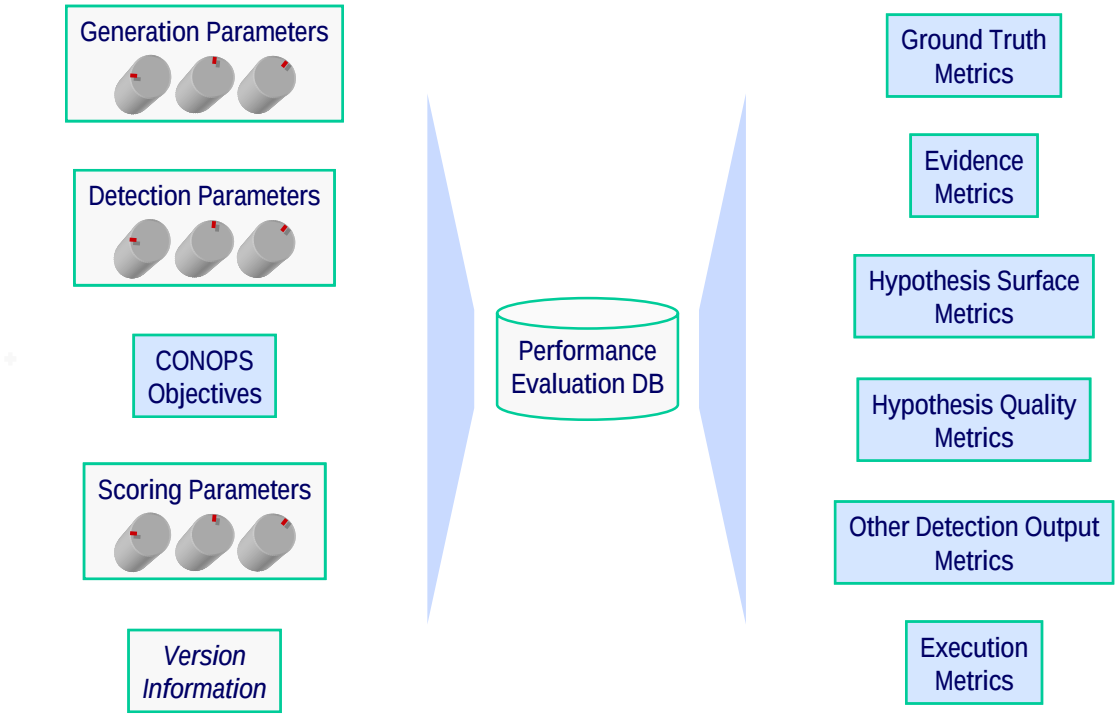

Figure 9: Historical performance evaluation information

Two kinds of elements to be stored in the performance evaluation DB make their first appearance in Figure 9:[2]

- Version information must be recorded for all PE Lab and Tangram system components, for each trial.
- "Other" detection output metrics result from processes (not shown in Figure 8) to characterize detection outputs other than hypotheses. These might cover numbers of matches to given patterns, as was explored during Tangram seedling experiments.

IET tenders that workflow processes producing evaluable hypothesis will have greater overall intelligence relevance and psychological validity to analysts than finer decompositions addressing pattern matching mechanics (where workflow processing reduces essentially to relational database query processing). The "blackboard"

[2] Like ground truth, generation parameters, while not tenable in operational contexts, may afford valuable diagnostic information in the development context.

architectural framework, schematized in Figure 10 to present the BAH SEA team's Tangram perspective, conceptualizes dynamic, hypothesis-oriented workflow.

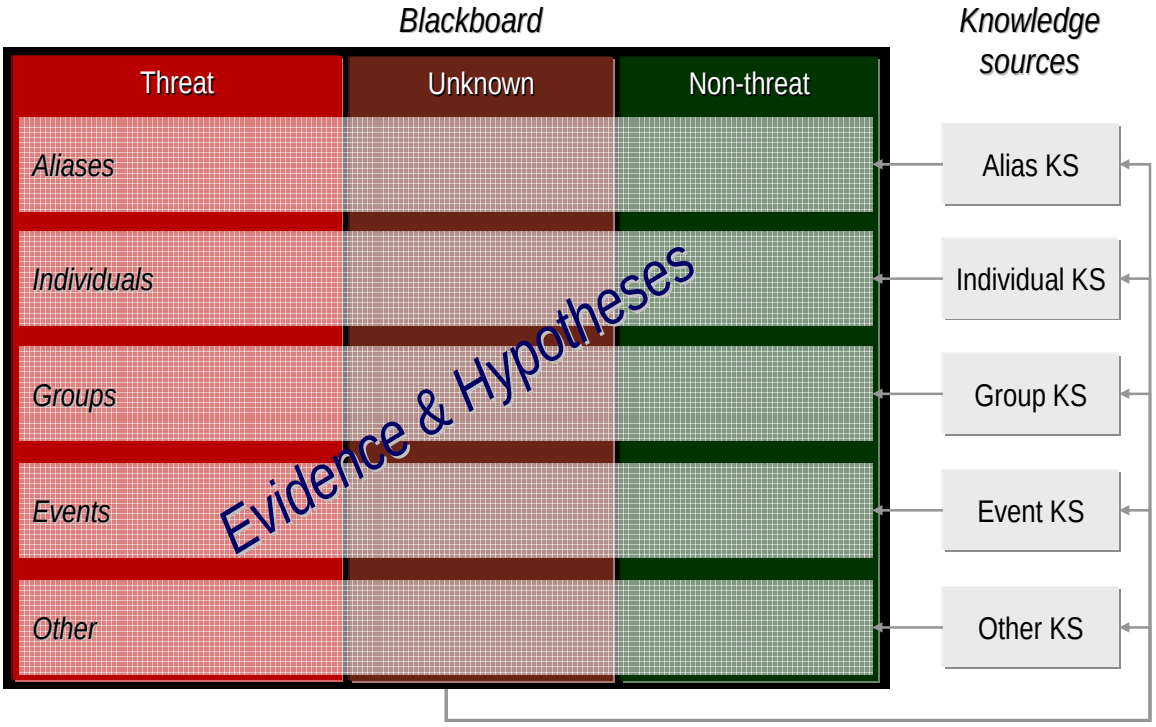

Figure 10: Blackboard perspective on Tangram workflow

The blackboard is a working repository of hypotheses.[3] Knowledge sources (KSs) process evidence and blackboard-resident hypotheses to develop further hypotheses.[4] Control knowledge (not shown in Figure 10) determines:

- Which hypotheses to retain and which to prune.

[3] Evidence is included here to render it accessible (schematically) to knowledge sources.

[4] EAGLE technology integration experiment (TIE) confederations participating in annual performance evaluations self-organized along lines of the top four KSs depicted in Figure 10—corresponding to the EAGLE PE Lab's scored hypothesis types. Different TIEs realized different degrees of KS cooperation in hypothesis production. In X-TIE, during the 2004 evaluation, USC/ISI delivered threat group hypotheses to Metron, who were then able to develop prospectively tendered threat hypotheses (AKA "alerts") much more effectively.

- When to invoke which KS, with what hypothesis, evidence, and other KS parameter inputs, to develop a workflow that is effective in addressing operational intelligence requirements.
- When to deliver extant blackboard-resident hypotheses for user consumption.

Fundamental to each of these decisions is the predictive assessment of hypothesis quality that PE Lab-based experimentation can inform.[5]

Figure 11 distinguishes among run-time and compile-time inputs available for each KS invocation, dynamic workflow development decision.

Figure 11 shows how the historical information covered in Figure 9 is compiled (using a generalization algorithm) to yield an induced performance model that complements another model that may be engineered directly from knowledge about the KSs' algorithms. The induced model may be used to check the engineered model's assumptions, or the induced model may actually replace the engineered one. The KS selection process applies its consensus model to the run-time inputs to select a KS for execution.

Figure 12 makes the Tangram blackboard's dynamic workflow explicit. The blackboard progresses through successive states as successive KS selections are made and KSs invoked.

This CONOPS can easily address the following variations on the dynamic workflow style we have presented:

- Static (compile-time) workflows
- End-to-end or piecemeal workflows
- Macro workflows that compose several individual workflow steps (as in a series of KS executions)

[5] Tangram systems also can employ the PE Lab to validate assumptions about the performance of a KS downstream, workflow-wise, from that of an upstream one. Suppose, *e.g.*, that a group KS depends on an alias KS to deliver sufficiently de-aliased evidence about individuals. If the alias KS is not yet performing at a goal level meeting the group KS's input specs, we can still ascertain validity of performance claims for the latter by stubbing the former with a direct feed of evidence having per-spec de-aliasing. This can help to pinpoint performance gaps among functional components early during development.

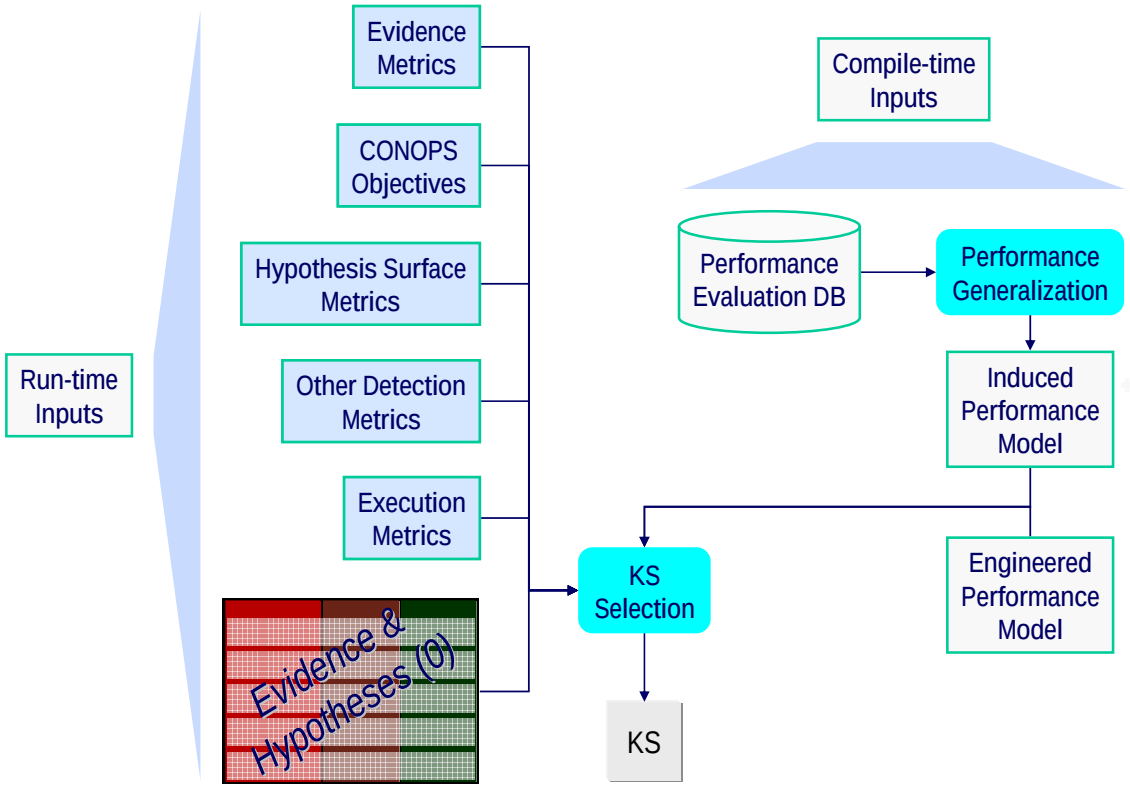

Figure 11: Dynamic workflow selection in the blackboard perspective (compile-time)

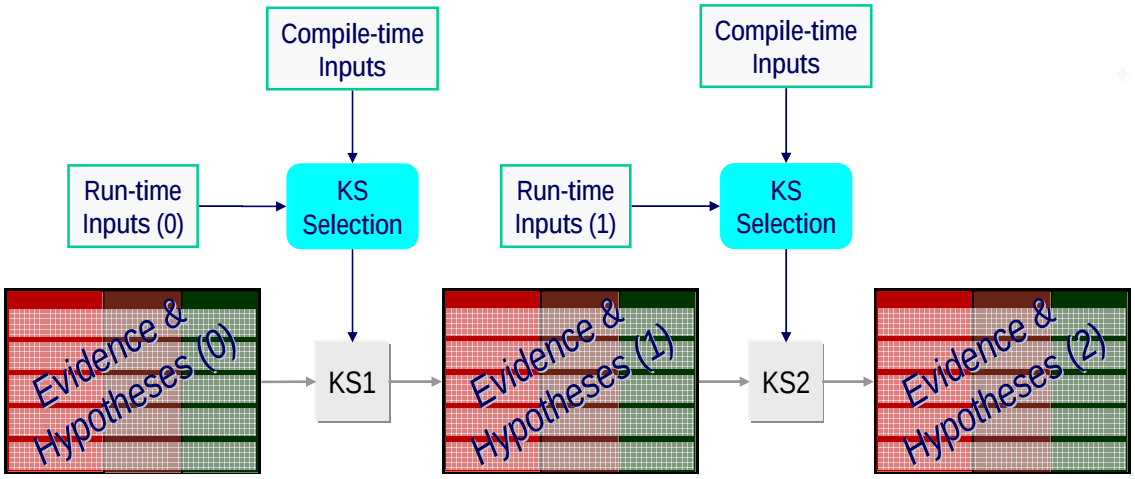

Figure 12: Dynamic workflow selection in the blackboard perspective (run-time)

6.2 Design of a Next-generation PE Lab

IET presents its recommendations for the design of a next-generation PE Lab to support future threat detection research and development. Subsections cover the following major topics:

- Next-generation PE Lab requirements (Section 6.2.1)
- Next-generation PE Lab design elements (Section 6.2.2)

6.2.1 Next-generation PE Lab Requirements

The next-generation PE Lab design uses dynamic social networks and explicit intelligence collection models to address requirements[6] summarized below:

- ***Provide a new synthetic data generator to produce unclassified datasets with the known characteristics of classified data sources.***
- ***Reflect the social networks that existing intelligence data sources portray.***
- ***Address groups that can add or lose members over time.***
- ***Address both the systematic characteristics of the intelligence collection process and the consequences of spotty collection and reporting of intelligence.***
- ***Address threat detection system feedback about existing information gaps in data collection***
- ***Explore the systematic characteristics of the intelligence collection process and our terrorist opponents to identify methods that will assuredly fail and methods that will produce the highest possible detection outcomes.***

Synthetic datasets are essential for evaluation of threat detection technologies because they:

- Provide ground truth against which threat detectors' hypotheses can be scored.
- Can be configured *via* key generation parameters to explore a wide range of detection challenges concerning factors like the following.
    - Observability in evidence of threat actors and their interactions
    - Masking in evidence of threat data by non-threat data
    - Similarity between threat and non-threat actors and interaction patterns
    - Temporal density of events

[6] These requirements are taken from the Tangram Proposer's Information Packet (PIP).

    - Overall volume and quality of evidence
- Enable focus on key detection challenges by abstracting away less relevant detail.
- Streamline access to information by directly generating structured data and employing a uniform ontology and associated database schema(s).
- Incur few issues of privacy or security classification associated with corresponding real-world datasets—thus are appropriate for experimentation by a wide research community.

The datasets generated using IET's EAGLE PE Lab have been used to:

- Assess detection technologies' progress against annual program performance goals.
- Identify characteristics of the detection problem that most influence a technology's performance.
- Predict which among available technologies will perform best on a given detection task.

IET's next-generation PE Lab design confers all the benefits listed above and meets requirements through use of dynamic social networks and explicit intelligence collection models. Specifically, the next-generation PE Lab design includes innovative capabilities to:

- Focus on dynamic social networks to maximize generality and ensure the harvest of low-hanging fruit.
- Develop social networks that reflect plausible states and interactions by:
    - Modeling growth and interaction patterns for threat and non-threat groups.
    - Developing networks from the ground up (starting from an empty world).
- Develop intelligence (for presentation as evidence to detection technologies) that is similar in character to real-world intelligence by:
    - Modeling intelligence collection processes explicitly.
    - Collecting intelligence throughout dynamic social network evolution.
- Support quantitative distinctions between threat and non-threat groups and their behaviors using parameters that may be varied continuously to determine detection technologies' relevant performance boundaries.
- Support (dynamic social network) objects and attributes that change over time by furnishing appropriate:
    - Evidence and ground truth temporal database semantics.
    - Temporal hypothesis scoring methods.

A **synthetic data generator** that can be used effectively across a wide variety of threat detection application contexts—with the only customization required per installation being to select among the available settings for some perspicuous, user-level parameters—must:

- Address detection objectives that are ubiquitous across the applications.
- Employ an ontology of entity types and relationships that is abstract enough to be universal across the applications while still affording an adequate basis of discrimination among threat and non-threat phenomena.[7]

In our performance evaluations for the EAGLE program, IET has used the PE Lab to pose challenges and the EAGLE Technology Integration Experiments (TIEs) have developed technology to address four different threat detection objectives—events, groups, individuals, and individual aliases. The EAGLE technologies pertaining to detection of threat actors (groups and individuals) have resulted in the greatest interest among and transition to analyst users in the Intelligence Community. Given this apparent leverage, IET has focused the next-generation PE Lab's design on the actor detection problem.

In the next-generation PE Lab design, social networks are developed starting with an empty set, using different timescales and different start times for expressing and generating the behaviors of individuals and groups, as illustrated in Figure 13.

This pervasively dynamic style ensures that social network evolution during the phase of evidence generation will be at least as realistic as that which generated the network in the first place. As such, the design meets the requirement to ***address groups that can add or lose members [or subgroups] over time.***

By specifying values (that might reflect either selected mappings from real-world datasets into the social network ontology or direct analyst intuitions) for the next-generation PE Lab's dataset generation parameters, the design meets the requirement that a user can ***produce unclassified datasets that reflect as closely as possible the known characteristics of classified data sources.***

Regarding the requirement to ***reflect the social networks that existing intelligence data sources portray,*** IET believes that the "holes" apparent in social networks from existing intelligence sources result from gaps inherent in the systematic intelligence collection process.

---

[7] Our initial ontology is described in Section 6.2.2.3.1.

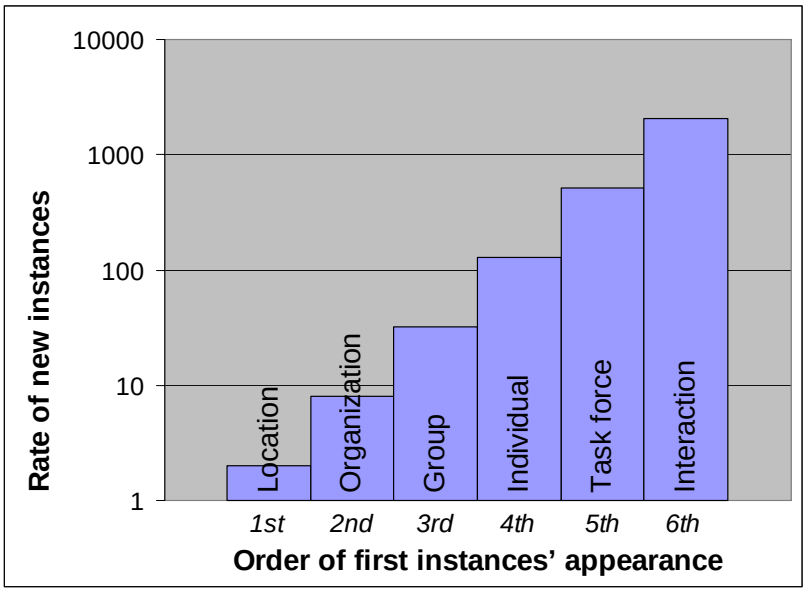

Figure 13: Time-phased introduction of simulation entities

The dataset generator in IET's EAGLE PE Lab develops evidence by uniformly applying—for each type of ground truth phenomenon that may be reported in evidence—a binomial distribution with a specified likelihood parameter. The next-generation PE Lab design meets the requirement to ***address both the systematic characteristics of the intelligence collection process and the consequences of spotty collection and reporting of intelligence*** by applying explicit intelligence collection models to ground truth phenomena to generate evidence. The next-generation PE Lab design models collection assets that are limited with respect to observed entities and (usually because of the way they are tasked) with respect to collection time intervals, yielding evidence that (while focused) can be sporadic and patchy. Modeled collection assets might collect the following.

- Communication transactions involving a specified individual or population sector
- Membership lists of cells or groups
- Transaction databases of communication providers

Note that the opportunities for collection focus don't necessarily mean that evidence will have a high signal-to-noise ratio (be rich in threat—*vice* non-threat—content)—it depends on what the asset observes. A collection asset has opportunities only randomly, according to probabilities associated with the asset and with evidence types. It also can exhibit corruption (resulting from collection error) and partial observability (resulting from collection incompleteness).

Figure 14 depicts a next-generation PE Lab architecture with taskable intelligence collection assets.

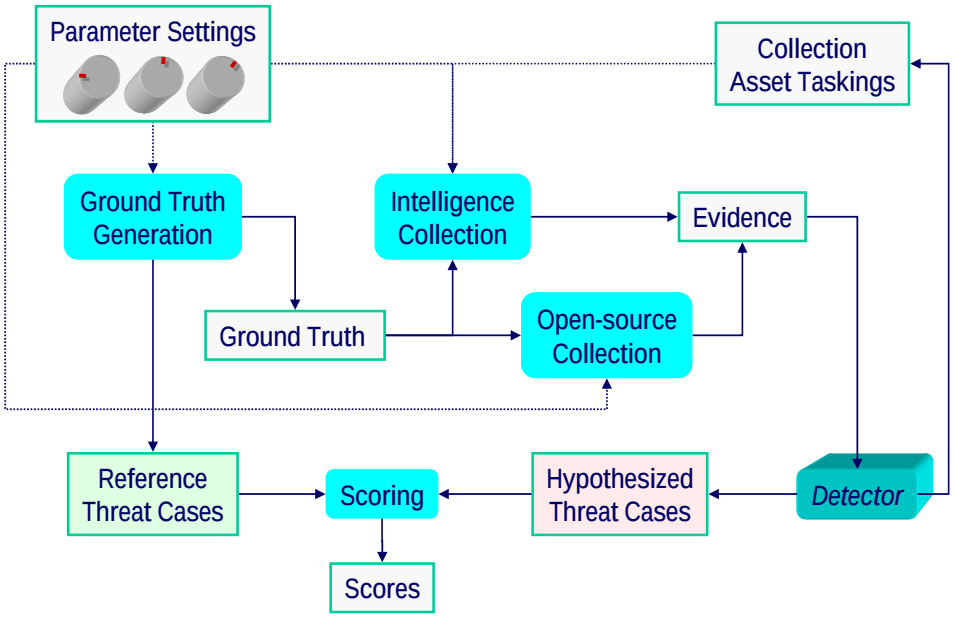

Figure 14: Evidence generation using tasked collection assets

In Figure 14, square-cornered boxes represent products/artifacts. Round-cornered boxes represent processes. The Detector or Link Discovery (LD) process (realized separately by each detection technology developer) is rendered 3-dimensionally to highlight its status outside of the PE Lab proper. Solid arrows represent flow of products/artifacts. Dotted arrows represent the flow of control information. Collection asset control is mixed. The Detector tasks an asset's attention, but the asset's success in representing extant phenomena accurately or completely will also depend on parameter settings that have been supplied for the dataset.

To maintain the unclassified status of resulting synthetic datasets, the design bases intelligence collection models on unclassified and open-source descriptions.

A next-generation PE Lab user (*e.g.*, an LD component) will be able to replay (using original dataset specifications and random seed) the generation of a given dataset and to invoke, from a simulation time point of interest going forward, an augmented collection model (that the user may modify at any succeeding time point) pertaining to specifically reserved user-taskable intelligence collection assets. We thus address the requirement to ***allow a detection system to provide feedback about existing information gaps in data collection.*** In simulation, all behaviors that pertain to collection will be independent of behaviors that do not, so that collection will have no effect on the ground truth.[8] (*E.g.*,

[8] The converse may not hold (i.*e.*, future collection targets may depend on past collection efforts.)

threat individuals will not be tipped off to collection methods so that they modify their behaviors; suspected planned attacks will not be prevented; suspected or known threat individuals will not be incarcerated and removed from the sphere of social interaction.) This approach is compatible with either incremental or batch processing of challenge datasets and enables performance comparisons across alternative technologies and/or collection policies for a given dataset.[9,10,11]

Some of the critical technology barriers faced by the next-generation PE Lab are familiar from the EAGLE context.

- **Scale:** EAGLE ultimately required datasets with 1,000,000 individuals conducting 10,000,000 observable transactions. To meet this requirement, IET has:
    - Streamlined PE Lab software to run much faster with much less memory.
    - Acquired faster hardware with more memory.
- **Variety:** The need for flexibility in experimentation drives synthetic dataset generation. It's not adequate to generate just one or a few datasets over an extended time period (say, months). Custom-generated datasets with a wide range of characteristics help developers to isolate and address weaknesses in their threat detection technologies. The tighter this loop is, the more quickly progress can be achieved. "Blind" datasets (for which answer keys and ground truth are withheld) used in one evaluation can't be used again for the next (after results have been published).[12]

---

[9] It would not be technically difficult to provide an API by which threat detection systems could affect the course of an ongoing simulation. Such a style (if preferable) might support—in addition to active collection—the actions of a larger "enforcement" (*vice* only detection) system empowered to arrest and incarcerate suspected threat individuals. This style would be possible only with incremental threat detection and would not as readily support comparisons (since the ground truth, though starting from identical initial conditions, would develop differently across the different simulation runs).

[10] Intelligence collection must be consistently managed across any integration of components in a threat detection system or workflow. *E.g.*, separate batch processes for threat event and group detection that task collection assets independently cannot be consistently combined in a streams-and-pipes workflow (such as was used by X-TIE during the 2004 EAGLE evaluation). Some consistent coordination of active collection over the workflow is called for.

[11] During a "blind" evaluation in which LD should not have access to ground truth, LD should be allowed to employ no more than one discretionary collection policy over any challenge dataset. (Different collection policies are likely to collect different information, which we do not want LD to accrete.) This may be best controlled by administering such evaluations *via* an independent test harness.

[12] "Open" sample datasets can be reused indefinitely.

- **Realism:** The critical issue is how to meet the above requirements and still deliver datasets with high experimental value. To mitigate complexity, we:
    - Use abstractions that strip away inessential detail from the detection problem.
    - Use in simulation only extremely lightweight agents (individual actors) without:
        - Explicit situational awareness
        - Intentions, plans, or schedules/calendars
        - Inter-agent activity coordination
    - Conduct multi-actor activities *via* scripts that are richly parameterized and randomized over the behavior patterns of interest.

In EAGLE, we established the effectiveness of the above overall strategy and also learned some lessons to carry forward into the next generation PE Lab's design:

- Threat events (*i.e.,* attacks) are difficult to develop good abstractions for (in large part because the specifics of real-world attacks—*e.g.,* target, method, and date—tend to have few good indicators in real-world evidence). The next-generation PE Lab design includes a limited notion of attack intended primarily to reveal the threat status of implicated actors.
- Artificial social networks used in simulation must exhibit (*e.g.,* power-law, scale-free, small-worlds) properties like those in the real world. We adopt the evolutionary and dynamic approach.

Given these observations, we see the following technical challenges as critical to the next-generation PE Lab's success:

- **Manage different timescales for introduction of simulation entities to streamline social network evolution while maintaining quality of evolved networks.**

  We believe simulation entities of different kinds should be introduced at different rates (as illustrated in Figure 13) so that social networks will evolve naturally. We might streamline evolution by limiting the amount of fine-timescale (*e.g.,* person-to-person interaction) behavior that takes place in a dataset's "pre-history" that precedes its portion that is presented as "contemporary" for threat detection. We don't want fine-timescale behavior completely shut down in the pre-history, because we rely on personal interactions to connect people socially and as opportunities for intelligence collection. Partial shut-down might require corresponding adjustments (to be rolled back upon commencement of the contemporary period) to the rates of connection and collection per interaction to maintain desired levels.
- **Manage parameters of social network evolution to realize desired evolved social network properties (emergently).**

Social network properties in the EAGLE PE Lab were largely realized directly *via* specified input parameter settings pertaining to (*e.g.*) the following:

- Number of individual actors
- Ratio of threat to non-threat individuals
- Sizes of threat and non-threat groups
- Ratio between numbers of threat and non-threat groups
- Numbers of groups that individuals belong to

In the next-generation PE Lab's evolutionary style, organizations[13] will acquire members *via* preferential attachment, and organization sizes will follow a power-law distribution. We can control the threat:non-threat ratio for organizations and for individuals by creating threat and non-threat organizations at different rates and by introducing individuals into them different rates. The number of organizations an individual belongs to should be consistent with the number organizations existing at a given stage of the network's evolution,[14] so should tend (on average) to grow as the network does. As noted above, we want individuals sometimes to join organizations following natural interactions with their members. We can limit the number of such joining events or wait for individuals to drop membership in some organizations before picking up others that might put them over their (individually) assigned limits.

- **Manage parameters of (intelligence and open-source) evidence collection to realize desired evolved evidence properties (emergently).**

  As noted above, the EAGLE PE Lab made individual ground truth phenomena observable according to probabilities that were uniform by phenomenon type. Evidence generation took place once, dataset-wide. In the next-generation PE Lab, where we model natural open-source and intelligence collection processes, simulated sources will (serially over time) issue reports about phenomena extant near their reporting times, and overall (*i.e.*, dataset-wide) observability at a given point in simulation time will depend on the following:

[13] EAGLE PE Lab "groups" lacked explicit subgroup structure. Next-generation PE Lab "organizations" will be composed of groups with (recursively) nested subgroups.

[14] It would be unnatural (*e.g.*) for every individual to belong to every group early in the simulation's pre-history. It would similarly be unnatural for an individual introduced in the simulation's contemporary period to join several organizations simultaneously.

- The number of collection assets active at any given (earlier or present) simulation time[15]
- The frequency with which each source reports
- The coverage of each source in each report and the method by which it moves from one reporting topic to the next[16]
- The rate at which ground truth phenomena (*e.g.*, the members of each threat group, the set of all threat groups) change[17]
- Any intrinsic observability properties associated with phenomena (*e.g*, actors' propensities to employ covert communication methods)

Given the above rich set of influences, we might achieve specified evidence observability (for a given type of phenomenon) in one of the following ways:

- Develop a closed-form analytical model that allows us to back into parameter settings so that sources can run as autonomous processes from simulation's outset to finish. This seems unnecessarily challenging and perhaps (given complex non-linear system dynamics) infeasible.
- Monitor evidence observability levels during simulation and adjust parameter settings to increase or decrease collection to meet observability targets. This entails the same requirement for analytical modeling—it remains difficult to predict the future effects of present asset introduction.
- Develop initial, candidate evidence using somewhat liberal parameter settings. Prune back excess evidence by cutting sources, reports, or per-report coverage (in ways consistent with the explicit collection models) proportional to their density at a given simulation time. If we could figure out how not to overshoot desired evidence levels by too much (again, potentially difficult), this method might be the most practical.

---

[15] It seems natural to introduce additional collection assets as social networks (and their underlying populations) grow.

[16] This will be different for different kinds of sources. A non-threat (*e.g.*, Government group) may publish only its own membership list. A COMINT wiretap specialist may move from one individual of interest to another following a chain of contacts.

[17] Descriptions collected in simulation pre-history might be obsolete during the era contemporary for LD's processing. In computing a dataset's observability at a given point in simulation time, we will consider only phenomena that have remained stable since their latest evidence was issued.

- Introduce sources in separate passes over the same (either stored or regenerated) simulation ground truth, collecting evidence incrementally until the overall dataset observability specified has been achieved. Multiple simulation runs might be prohibitively expensive if many were needed.
- Develop, for each observed phenomenon type, a regression model regarding overall dataset observability as a function of intelligence collection and other relevant dataset generation parameters. This method might practically be used (as a surrogate for analysis) in combination with one of the two preceding methods.

We feel that explicitly modeled intelligence collection assets will yield benefits to offset any potential difficulties with controlling the observability of threat phenomena in evidence—because it will afford the experimental basis to address the requirement to ***explore the systematic characteristics of the intelligence collection process and our terrorist opponents to identify methods that will assuredly fail and methods that will produce the highest possible detection outcomes.***

6.2.2 Next-generation PE Lab Design Elements

Subsections cover the following key design foci:

- Dynamic social networks (Section 6.2.2.1)
- Explicitly modeled intelligence collection (Section 6.2.2.2)
- Supporting foundations (Section 6.2.2.3)

6.2.2.1 Dynamic Social Networks

Real-world social networks tend to exhibit scale-free, power-law phenomena, as depicted in Figure 15.

Figure 15 pertains to an artificial network of 800 persons generated[18] using the procedure listed below.

For the number of individuals desired:

1. Introduce a new individual *I.*

[18] IET's experiments to generate this data were inspired by the body results of summarized in Barabási, Albert-László. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*, Plume, 2003.

2. **Preferential attachment:** Connect *I* (as an acquaintance) to a different, existing individual *D*, whose choice is weighted by current degree.
3. For the number of triangulations desired per new individual:
   a. Select an existing individual *A* at random.
   b. **Triangulation:** Connect (as acquaintances of each other) two existing acquaintances *B* and *C* (randomly chosen) of *A*.

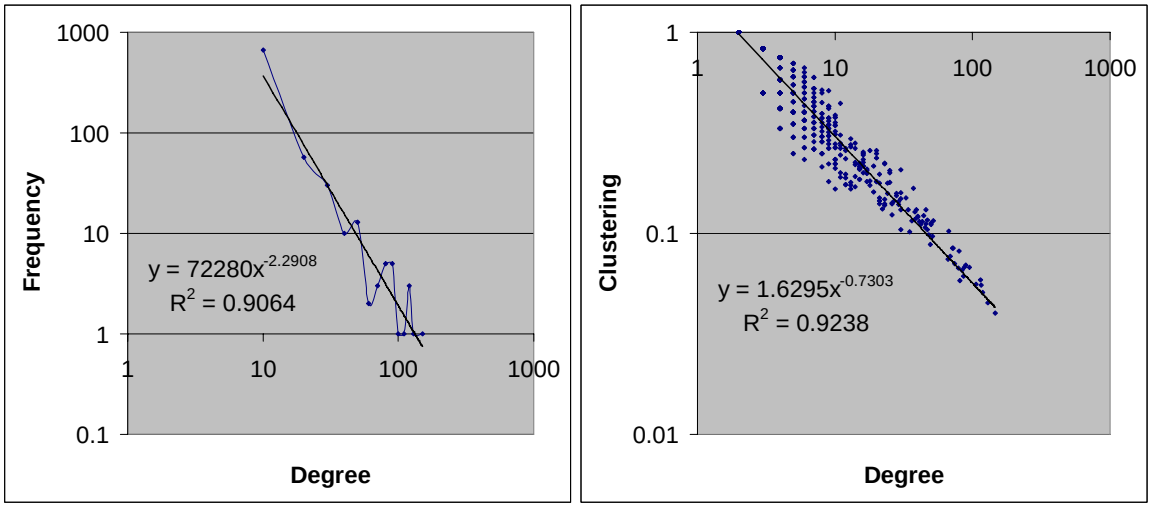

Figure 15: Properties of a scale-free network—degree distribution (left) and clustering coefficient[19] by degree (right)

Preferential attachment and triangulation lead to the power-law distributions in Figure 15. People who know more people tend to get introduced to more people in more remote groups.[20] Through experimentation, IET has learned that both preferential attachment and triangulation are needed for the best fits to power-law equations. Of the two, triangulation appears to have the greatest influence.

We want to exploit these phenomena in the next-generation PE Lab's dynamic social networks, but we must move beyond the above simple model to address structured

[19] The clustering coefficient of a node *N* is defined as *a(N) / p(N)*, where *a(N)* is the number of acquaintances of *N* who are acquainted with each other and *p(N)* is the maximum possible value for *a(N)*.

[20] Also (not shown in Figure 15, but true of its network), shortest path lengths between any two people tend to be small.

organizations. As a point of departure, consider the hierarchical, cell-based threat organization depicted in Figure 16.

The organization includes seven groups (the cells), each of which (notionally) includes five members. The top cell is the leadership cell, and the top member in each cell is the cell leader. Only one member in any cell is acquainted with any member in any other cell, so the leaders of subordinate cells must be different from the cell leader. Interactions within each cell are highly connected.[21]

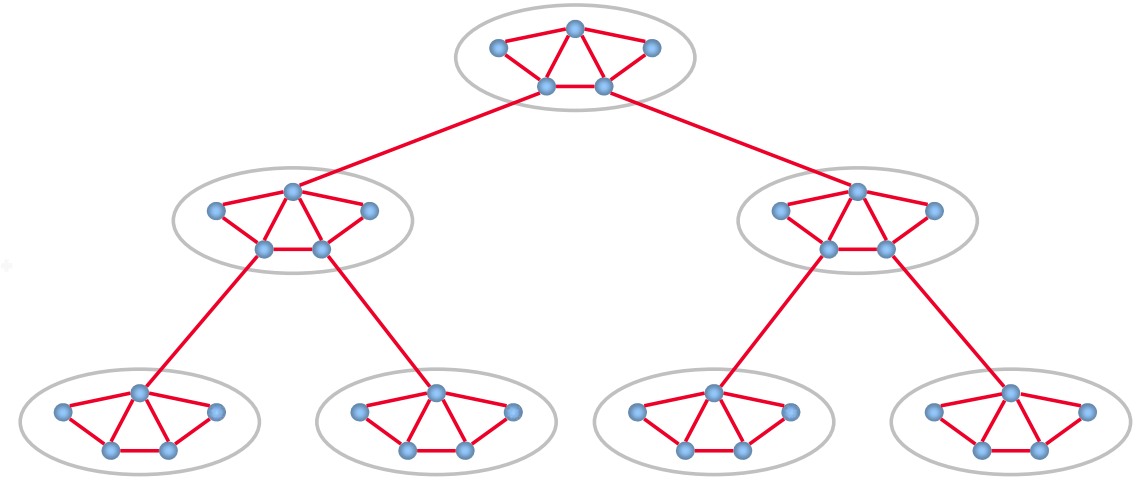

Figure 16: A hierarchical, cell-based organization

Such a cell-based organization is anomalous compared to the typical sort of social network profiled in Figure 15 in that individuals' degrees and clustering coefficients vary little over the network. Shortest path lengths also tend to be longer than in a normal network. The purely cell-based organization also presents the following difficulties:

- Cells (and subcells) once disconnected because of the loss of a leader or his controlling outside contact cannot easily reconnect to the organization.
- Cells can never split or merge to form new cells.
- Presuming limits on numbers of members in a cell and on numbers of subcells an individual can control, most organization growth must occur around leaf cells. Unless candidates connect here directly, they may not succeed in joining.

[21] We have refrained from showing the fully connected group interaction graphs.

Cell-based organizations can (and typically do) overcome these difficulties by relaxing the pure cell doctrine. Some potential relaxations are show in Figure 17.

Figure 17 uses light gray background shading to depict some interactions and connections among (shading-surrounded) individuals that breach the pure cell doctrine. The more the cell doctrine is relaxed, the more a threat organization resembles (in character) a non-threat organization.

The next-generation PE Lab design parameterizes a general notion of hierarchy to realize dynamic organizations covering a variety of styles. Table 1 suggests how parameters affecting organization structure, growth, and interaction can be set to realize some different styles of interest.

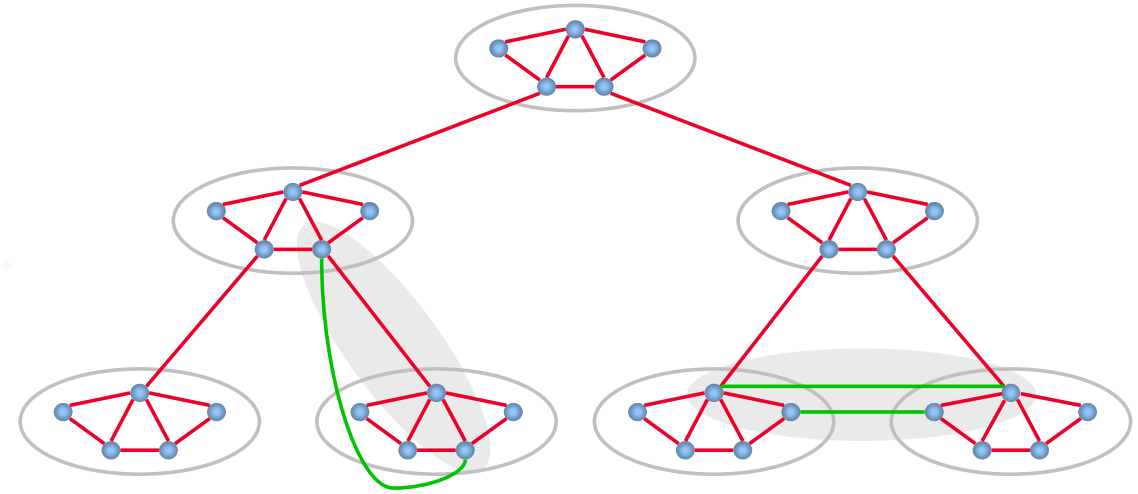

Figure 17: Relaxing the cell doctrine

	Purely cell-based	*Loosely cell-based*	*Normal hierarchical*	*Loosely flat*	*Purely flat*
Members per group	[2, 12]	[2, 12]	[2, 24]	[2, ∞]	[2, ∞]
Task forces per group	1	1	[1, ∞]	[1, ∞]	[1, ∞]
Primary task initiator	Manager	Either	Either	Member	Member
Multi-group task forces	Never	Rarely	Occasionally	Frequently	n / a
Subgroups per group	[0, 10]	[0, 10]	[0, 10]	[0, 2]	n / a
Managers per group	[0, 10]	[0, 10]	[0, 10]	[0, 2]	0
Subgroups per manager	[1, 5]	[1, 5]	[1, 5]	[0, 2]	n / a
Subgroup growth control	Low	Low	High	High	∞
Task force reviews	Frequently	Frequently	Frequently	Rarely	n / a
Cross-echelon reviews	Never	Rarely	Occasionally	Rarely	n / a
Group mergers	Never	Rarely	Occasionally	Rarely	n / a
Group splits	Never	Rarely	Occasionally	Rarely	n / a

Table 1: Parameters conferring different styles on generated organizations

Note the following regarding the concepts included in Table 1:

- **Groups** (of individuals) are structural elements of organizations. Groups in threat organizations are **cells.**
- The **normal hierarchical**[22] style is typical of business organizations. We expect power-law social network phenomena in these organizations because members at higher echelons tend to have more contacts with individuals outside their immediate working groups. Managers meet with subordinates in their own organizations. Higher-echelon managers have (presumably shorter) more frequent meetings. Lower-echelon members gain access to higher-echelon ones at different rates. Triangulation is at work here, as managers refer high-value messengers up the management chain. Managers also meet with peers in other organizations and perhaps develop organization-wide relationships.
- **Task forces** are collections of individuals cooperating to perform particular tasks.[23] Task forces may be permanent or temporary. Subsets of their members may interact frequently or infrequently towards the task's performance. Task force members usually all belong to the same group but also may be taken from

[22] The **normal matrix** style (not shown in Table 1) differs from the **normal hierarchical** style only in its greater frequency of multi-group task forces.

[23] Our initial challenge problem ontology in Section 6.2.2.3.1 includes no notion of task, as we focus on task force interactions.

different groups (as in **multi-group task forces**)—typically one nearby in the same organizational hierarchy but possibly from a different organization.

- Task forces may periodically meet with their managers for **task force reviews** that may subsequently be escalated to **cross-echelon reviews** with higher management levels.
- Individuals generally will belong to multiple organizations. For simplicity, an individual will belong to only a single group of a given organization. Individuals will join their first organization *via* preferential attachment, subsequent organizations *via* triangulation.
- The **loosely flat** style is typical of a socially (*vice* economically) oriented organization (for, *e.g.*, recreational sports, community service, or religion) with relatively little structure. (The **purely flat** style, reflecting an "organization" with no structure, is included for sake of comparison, not necessarily to be realized in simulation.)
- Tasks in social organizations are typically **initiated** by members. Tasks in economic or threat groups are typically initiated by either members or managers.
- In economic or threat organizations with small group sizes, we assume that all group members are acquainted with each other and may serve together on any task force. In social organizations with large group sizes, the members of new task forces are linked by existing acquaintance relationships (in current task forces).[24]
- A group's **subgroup growth control** refers to:
    - The number of members it must have before it may spawn subgroups.
    - The number of subgroups it must have before its subgroups may spawn subgroups.

6.2.2.2 Explicitly Modeled Intelligence Collection

Different sources collect different types of intelligence at different rates. Figure 18 describes coverage of phenomena of interest by collection type.

[24] We can't practically store all the individuals another has ever interacted with. We may also need to limit the number of task forces an individual may belong to.

Interaction type \ Source	Focused surveillance	Broad surveillance	Connectivity provider	Open source
Unidirectional two-party				
Letter	■	■	■	
Email	■	■	■	
Word-of-mouth courier	■			
Unidirectional multi-party				
Open media (e.g., Web)				■
Bidirectional two-party				
Phone	■	■	■	
Instant message	■	■	■	
In-person meeting	■			
Multidirectional multi-party				
Teleconference	■	■	■	
Chat room	■	■	■	
In-person meeting	■			

Group property \ Source	Open source	Forensics	Informant (member)	Tipster
Threat group membership			■	■
Non-threat group membership	■			
Attack agency	■	■	■	

Figure 18: Coverage of phenomena of interest by collection type

We note the following regarding Figure 18:

- Broad surveillance (always COMINT) does not cover in-person interactions. It may cover limited geographical regions, and it may cover different regions at different times.
- Focused surveillance (which may be *via* either HUMINT or COMINT) usually covers a more specific target than broad surveillance.
- Connectivity providers may serve limited population segments.
- The evidence available regarding remote interactions won't necessarily identify the interactors unambiguously. Communication termini such as phones, addresses, and user IDs can be shared by multiple individuals (some of whom act deliberately to deceive threat detection systems).
- Attacks may serve to reveal the existence of threat groups/organizations and some of their membership.

- The next-generation PE Lab design models HUMINT and COMINT sources with different coverage and accuracy rates regarding different phenomena of interest.[25]
- The next-generation PE Lab design routinely initiates some focused intelligence collection against individuals observed to interact with or connected by forensics to other, suspected threat individuals.
- Non-threat organizations may publish some of their top-level membership and organizational structure on a regular basis. Informants may provide more complete information (*e.g.*, a company or division phone directory).

6.2.2.3 Supporting Foundations

Subsections introduce the following supporting foundations for dynamic social networks with explicit intelligence collection models:

- Initial challenge problem ontology (Section 6.2.2.3.1)
- Scored hypothesis types (Section 6.2.2.3.2)
- Representation requirements for the database schemas supporting ground truth and evidence (Section 6.2.2.3.3)
- Temporal representation and reasoning for scoring (Section 6.2.2.3.4)

6.2.2.3.1 Initial Challenge Problem Ontology

Below is the initial challenge problem ontology to be used in evidence. (Internal and/or ground truth versions may have additional content for simulation support.) A **Class** appears (perhaps multiple times) with indentation to indicate its position in the class lattice. A *classAttribute* falls directly under its defining **Class** and is followed by the (Class) of its single value or (Class*) of its multiple values. For example:

Actor
Individual
location (Location)
aliases (Individual *)
communicationTermini (CommunicationTerminus*)
ThreatIndividual
NonThreatIndividual
Group

[25] Rates will generally be under parameter control. Realistic rates (key to assessing a technology's effectiveness, but not required in all experiments) will be elicited from an experienced counter-terrorism intelligence analyst.

leader (Individual)
directSubGroups (Group*)
directSuperGroup (Group)
directMembers (Individual*)
organization (Organization)
ThreatGroup
ThreatOrganization
NonThreatGroup
NonThreatOrganization
Organization
ThreatOrganization
NonThreatOrganization
Source[26]
Event
startDate (Date)
endDate (Date)
Interaction[27]
participants (Individual*)
Attack
participants (Individual*)
perpetrator (Group)
Location (Integer)
CommunicationTerminus[28]
communicationProvider (CommunicationProvider)
CommunicationProvider (Integer)
Date (Integer)

6.2.2.3.2 Scored Hypothesis Types

The next-generation PE Lab design scores the following LD-output hypothesis types, using methods developed in the EAGLE PE Lab:

- Threat individuals and their aliases and communication termini
- Threat cells[29] and their member individuals

[26] See Figure 18 for potential refinements of this class.

[27] See Figure 18 for potential refinements of this class.

[28] See Figure 18 for potential refinements of this class.

- Threat organizations and their (indirectly) constituent cells

6.2.2.3.3 Representation Requirements for the Database Schemas Supporting Ground Truth and Evidence

Dynamic social networks require temporal representation and reasoning. Evidence may be supported by explicit probabilities regarding a source's general credibility on certain kinds of evidence or may include explicit probabilities regarding a source's professed certainty about particular asserted propositions. We describe feasible temporal and uncertainty representations and associated schema requirements for a shredded (third-normal-form) relational database.[30]

More expressive representations—supporting more kinds of ambiguity and/or uncertainty regarding world states—require more complex reasoning by LD. IET's bias is towards an expressive schema that will support evidence reasoning challenges across a range of difficulty (making the complexity of evidence another potential dimension of the LD problem space).

Subsections address these topics:

- Temporal requirements (Section 6.2.2.3.3.1)
- Uncertainty requirements (Section 6.2.2.3.3.2)

6.2.2.3.3.1 Temporal Representation Requirements for Ground Truth and Evidence

Ground truth, being unambiguous, affords a point of departure regarding temporal representation requirements.[31] We need to note the points in simulation time at which propositions pertaining to objects' existence and to their attributes' values begin to be in effect and the points at which they cease to be in effect. Consider a group (with UID) Gr-123 that is formed on simulation Day 12, is (from the outset) part of organization Or-246,

---

[29] We include no hypotheses for task forces because (per Table 1) each threat group cell corresponds identically to a task force.

[30] See Raghu Ramakrishnan's (U. Wisc.) October, 2006 "EDB 2" Tangram working group presentation, to which IET contributed.

[31] Depending on the economics of generating a given dataset, it may even be advantageous to materialize elements of ground truth (whose volume may be much greater than that of evidence) only on demand *via* dataset regeneration (from a stored random seed) rather than actually to store ground truth. We would nonetheless materialize and store the "answer key" portion of ground truth pertaining to threat phenomena.

and has (for a time) among its members individual In-456. In-456 moves from location Lo-2 to location Lo-7. Our ground truth schema would include records in tables with fields[32] as shown in Figure 19.

Group table fields:	groupUID	observationTime
Group record values:	Gr-123	(Day 12)

Group_organization table fields:	groupUID	observationTime	organizationUID
Group_organization record values:	Gr-123	(Day 12)	Or-246

Group_memberAgents table fields:	groupUID	observationTime	personUID
Group_memberAgents record values:	Gr-123	(Day 14)	In-456

Group_notMemberAgents table fields:	groupUID	observationTime	personUID
Group_notMemberAgents record values:	Gr-123	(Day 17)	In-456

Person table fields:	personUID	observationTime
Person record values:	In-456	(Day 12)

Figure 19: Ground truth schema example

Note the following:

- Separate tables pertain to an object's existence and to the values of its attributes.
- Each attribute table covers just one attribute, so that records for different attributes can cover different time intervals. Such tables also conveniently handle multiple contemporaneous values of the same attribute (like multiple members of a given group).
- Each table includes a field (here, groupUID) to designate the subject object and a field (observationTime) to designate the simulation time at which the object is observed. In a ground truth database, the latter is always a simulation time at which the subject object acquires the subject property.
- The kind of table used to record an object's state changes differs depending on the semantics of the property.

[32] Only fields relevant to the example are shown.

- For single-valued attributes (*e.g.*, Person_location) for which an object always must have some value, we simply record the later value with onset time in a separate record.
- For other single-valued attributes (*e.g.*, Group_organization) and for multi-valued attributes, (*e.g.*, Group_memberAgents), we must assert the explicit negation of the attribute value. We use explicitly negating tables (*e.g.*, not_Group_memberAgents) with the same fields used in the non-negating tables.

The observations reflected in evidence—unlike those in ground truth—do not constitute a comprehensive description of the world across simulation time. Rather than a single, omniscient source, we may have multiple, imperfect sources that may report different versions of the same object's state at a given simulation time.[33] If we suppose each source observes (aspects of) an object at a point in time, we can accommodate reports pertaining to these "observation intervals" by augmenting the tables of the ground truth schema above[34] with "report" fields, as illustrated in Figure 20.

Person_location table fields:	personUID	observationTime	locationUID	reportUID
Person_location record values:	In-456	(Day 15)	Lo-2	Re-7890

Report table fields:	reportUID	reportTime	sourceUID
Report record values:	Re-7890	(Day 15)	So-2

Figure 20: Evidence schema example

The report time is the simulation date of the report's issue by the source and the date at which an LD component, processing evidence incrementally as it steps through simulation time, may examine the report's evidence.[35]

---

[33] The EAGLE evaluation DB schema eschews the notion of report (multiple or otherwise), with the primary intents of streamlining the schema and reducing overall data volume. While that approach remains an option going forward, it would not (realistically) challenge LD to reason about contradictory reports, including those that may arrive at different simulation times but pertain to the same simulation time.

[34] Note that we continue to use definite time points for observations. Indefinite observation points—*e.g.*, bounded by the constraint (contains [(Day 911) (Day 999)] ?startDate)—require additional tables for the (abstract) time points that are indefinite and for constraints on these. Modeling some kinds of real-world evidence may require such indefiniteness.

[35] LD must request reports issued since its similar request at a prior incremental time step before it may examine the evidence records designated as being contained in each such report.

Suppose evidence contains the observation that individual In-456 was a member of group Gr-123 on simulation Day 23.[36] What may a consumer of evidence (such as an LD component) believe about In-456's membership before or after that time? In temporal logic terms, should the assertion of his membership be assumed to "persist" backward (or forward) in simulation time from the observation point, absent an earlier (or later) contradictory report? Our ground truth schema relies on forward persistence. Backward persistence also is needed in evidence, where we can't assume that a first time of observation corresponds to an object's time of state change. If it happens so to correspond, the report can include an explicitly negating observation at the immediately preceding simulation time point—if it does not already include a contradiction of some other kind pertaining to that time. Potential varieties of contradiction include:

- A functional dependency declared in the domain ontology (*e.g.*, "He can't be in locations Lo-2 and Lo-7 at the same time.")
- The limited temporal extent of an object's existence (*e.g.*, "He died on Day 38.")

While we can provide an adequate representation, an assumption of persistence (like any evidence) is really left to the consumer to tender belief in or not. A consumer may wish to consider the following issues:

- It may be appropriate to limit belief (or degree of belief) as a function of temporal distance from the observation point.
- Any gap in time between two contradictory observations (absent other relevant, intervening assertions) leads to ambiguity regarding where during the gap the observed object changes state.
- A later contradictory report (that may pertain to a relatively early simulation time) may present a requirement for belief maintenance. The consumer should be prepared to withdraw belief in the assumed proposition (and perhaps in any assumption-based conclusions it has derived).
- Uncertainty of the following varieties may impinge on the persistence assumption:
    - A report's source may have imperfect credibility.
    - A report's source may specify uncertainty regarding the propositional content of its observation.

[36] The actual time granularity of simulation may be finer or coarser than daily.

- The persistence may emanate from an indefinite observation time point. (A report's source may specify uncertainty regarding the temporal content of its observation.)

6.2.2.3.3.2 Uncertainty Representation Requirements for Evidence

Evidence may be supported by (supplied) models including explicit probabilities regarding a source's general credibility on certain kinds of evidence. Evidence from a given source may include explicit probabilities regarding its professed certainty about particular asserted propositions. To support credibility reasoning, a (non-temporal) table pertaining to sources may report (or may not report, depending on the observability of a given source's credibility model) the probability that the source will corrupt a specified field of a given record in an object existence or object attribute value table (*e.g.*, in the table Group_memberAgents, groupUID, observationTime, or even professed proposition certainty).[37] To support professed assertion certainty, we add a probability-valued "certainty" field to the object existence and object attribute value tables of the evidence schema, as illustrated in Figure 21.

Group table fields:	groupUID	observationTime	certainty
Group record values:	Gr-123	(Day 12)	0.95

Figure 21: Example evidence schema with propositional uncertainty

We prefer to keep the temporal and uncertainty aspects of representation separate, where possible—so that we can vary their expressiveness and the complexity of LD's reasoning challenges independently, if desired. We also prefer (initially) not to clutter the evidence schema described so far by supporting arbitrary expressivity for uncertainty. Therefore, we avoid explicit (source-professed) conditional dependencies among the assertions for a given object (or even across objects) at a given time (or even across times) in a given report (or even across reports). We have the following conditional dependencies implicitly:

- Actual object existence probability depends on professed object existence probability and on source credibility pertaining to the latter.
- Actual object attribute value probability depends on professed object attribute value probability and on source credibility regarding pertaining to the latter.

[37] We generally can arrange for these probabilities to be independent without compromising realism too much.

- By convention, any two contradictory observations that are contained in the same report and have adjacent simulation times also will have the same professed probability, which may be interpreted as applying to their conjunction.

6.2.2.3.4 Temporal Representation and Reasoning for Scoring

Dynamic social networks also present temporal representation and reasoning requirements to support scoring of LD-output hypotheses. Subsections address the following topics:

- Requirements for answer keys (Section 6.2.2.3.4.1)
- Requirements for evidence summaries (supporting relative scoring—Section 6.2.2.3.4.2)

6.2.2.3.4.1 Temporal Representation and Reasoning for Answer Keys

The next-generation PE Lab dynamic social networks require a temporal representation for answer keys, so that hypotheses submitted at any simulation time can be scored with respect to the ground truth at that time. The required information is contained, for each scored hypothesis type, in the corresponding class' ground truth scored attribute tables. The ground truth state of a given scored class instance at a given simulation time is determined by the forward-persistent semantics of the ground truth representation.[38]

Presuming LD returns its hypotheses in a like format, we can develop both ground truth's and LD-output hypotheses' versions of the set of all scored objects at any given point in simulation time, then score these using existing EAGLE methods that apply to static situations.[39] This reuse of the ground truth schema for LD-output hypotheses presumes (as in EAGLE) unambiguous hypotheses. We would need to develop a separate hypothesis schema with associated reasoning methods should we desire to accommodate in scoring ambiguous hypotheses employing indefinite time points or specifying non-unit probabilities for the properties of scored objects. As noted in Section 6.2.2.3.4.2, we need a capability along these lines for datasets with similarly ambiguous evidence to support EAGLE's relative scoring methods.

6.2.2.3.4.2 Temporal Representation and Reasoning for Evidence Summaries

Intelligence analysis regarding standing types of threats is a continuous process. Analysts routinely consume raw intelligence (and open-source) reports and produce

---

[38] Take the latest value at or before the time of interest for any property of any object.

[39] Depending on the number of objects to be scored for a given dataset, we may be able practically to score hypotheses only at selected simulation time points. We need not, however, inform participants prior to hypothesis submission in a "blind" evaluation what the scoring time points for a given dataset will be.

summaries. While many such summaries are text reports intended primarily for decision makers, some are computer-manipulable models, such as counter-terrorism databases or link analysis diagrams, which can also serve as analytical resources. In the EAGLE PE Lab, LD is presented with a summary[40] of known evidence regarding threat actors and events, along with unprocessed evidence. LD's challenge is to update the summary (the way an analyst would maintain a model), given the unprocessed evidence, so that the summary reflects (current[41]) ground truth as accurately as possible.

The summary schema is like the hypothesis schema in that it lacks reports and describes each threat phenomenon at any given simulation time only once. It also has the following fields to support computation:

- Explicit probabilities (that may be taken to reflect the combination source credibility and professed assertion certainty)
- Explicit, independent backward and forward persistence assumptions (that may, for a given proposition, be dropped during computation of the summary)[42]

In the next-generation PE Lab, where (in contrast to EAGLE) different raw reports can describe the same or different threat actors at the same or different simulation times, we must compute such a summary from the reports. This summary, computed at the outset of the dataset's contemporary simulation era, provides important seed data[43] for threat detection algorithms. How we compute the input summary depends on what evidence about threat actors we consider to be "obvious." The following evidence representation capabilities pose issues for summarization:

- Source's credibility and reports' professed assertion certainties
- Temporally overlapping contradictory reports, including those arising from opposing persistence assumptions and those from indefinite time points

In general, we desire a summary whose contents all meet or exceed some threshold probability. We must investigate inexpensive ways to improve the soundness and completeness properties of the following, approximate method.

---

[40] In EAGLE, this summary includes the threat phenomena reported in "primary" evidence.

[41] A given real-world model might also include relevant historical (*i.e.,* non-current) information. We might augment our evidence summary schema to accommodate historical information. This would require corresponding augmentation of scoring methods.

[42] We have omitted these from the evidence schema for the sake of compactness.

[43] Only some of a dataset's threat phenomena are covered by seed data. The remainder technologies must detect without benefit of seeds.

1. Develop a probability-thresholded representation.
    a. Begin with a (potentially inconsistent) temporal representation reflecting the contents of all reports in evidence.
    b. Discard any contradictory assertions that have the same observation timepoint.
    c. Discard all assertions not meeting the threshold probability.
    d. Discard all assertions with indefinite observation timepoints.
    e. Treat all remaining assertions as if they were certain.
    f. Drop any opposing persistence assumptions for contradictory assertions.
2. Compute implied properties of threat object instances at the timepoint of interest.
    a. Compute threat groups/organizations.
        - Those explicitly mentioned as threat groups/organizations
        - Groups whose stated organization is a known threat organization
        - Groups for which a direct or indirect subgroup or supergroup relationship to a known threat group/organization can be determined from directSubgroup or directSupergroup assertions and/or from organization assertions
        - Groups stated to be perpetrators of threat events
    b. Compute threat individuals.
        - Members of threat groups/organizations
        - Participants in threat events

Another summary, computed at the end of the dataset's contemporary simulation era (or at any simulation time where scoring is desired), serves (in relative scoring[44]) as a basis for evaluating threat detection algorithms' updated, product summaries. Note that intelligence collection asset tasking by LD will result in a different (presumably richer) product summary than that provided for evidence routinely collected. Thus, in the next-generation PE Lab, the following varieties of relative scoring may be of interest.

- LD's hypotheses relative to routinely collected evidence summary—*What fraction did LD attain of the difference between a perfect score and that of the*

[44] Relative scoring compares the quality of LD's output summary to that of its input summary.

summary for evidence routinely collected? (Or, what benefit did LD's hypotheses confer relative to a straightforward summary of routinely collected evidence?)

- LD's hypotheses relative to routinely collected plus LD-collected evidence summary—*What fraction did LD attain of the difference between a perfect score and that of the summary for evidence routinely and/or actively collected? (Or, what benefit did LD's hypotheses confer relative to a straightforward summary of all collected evidence?)*
- Routinely collected plus LD-collected evidence summary relative to routinely collected evidence summary—*What benefit did LD's active collection confer relative to routine collection?*

7 REFERENCES

Barabási, Albert-László. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*, Plume, 2003.

8 LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

Table 2 serves as a reference for all symbols, acronyms, and abbreviations, along with their expansions, that are used and contained throughout this document.

Acronym, Symbols, or Abbreviation	Expansion
AFRL	Air Force Research Laboratory
AI	Artificial Intelligence
AKA	Also Known As
API	Application Programming Interface
BAH	Booz Allen & Hamilton
BN	Bayesian Network
COMINT	COMmunications INTelligence
CONOPS	Concept of Operations

CSV	Comma Separated Value
DARPA	Defense Advanced Research Projects Agency
DB	DataBase
DoD	Department of Defense
EAGLE	Evidence Assessment, Grouping, Linking, and Evaluation
HUMINT	HUMan INTelligence
IC	Intelligence Community
ID	IDentification
IET	Information Extraction & Transport, Inc.

ISI	Information Sciences Institute
KS	Knowledge Sources
LD	Link Discovery
PE Lab	Performance Evaluation Laboratory
PIP	Proposer's Information Packet
POC	Point of Contact
SEA	System Evaluation and Architecture
SPS	Software Product Specification
SUM	Software User's Manual
TIE	Technology Integration Experiment
UID	Unique Identifier
USAF	United States Air Force
USC	University of Southern California
Y3	Year 3
Y4	Year 4

Table 2: Acronyms / Expansions